EXHIBIT Q

TO DECLARATION OF S. MERRILL WEISS IN SUPPORT OF PLAINTIFF ACACIA MEDIA TECHNOLOGIES CORPORATION'S MEMORANDUM OF POINTS AND AUTHORITIES IN OPPOSITION TO ROUND 3 DEFENDANTS' MOTION FOR SUMMARY JUDGMENT OF INVALIDITY UNDER 35 U.S.C. § 112 OF THE '992, '863, AND '702 PATENTS; AND SATELLITE DEFENDANTS' MOTION FOR SUMMARY JUDGMENT OF INVALIDITY OF THE '992, '863, AND '720 PATENTS

Hovsepian v. Apple, Inc.
Doc. 325 Att. 17

Dockets.Justia.com

Transform Coding of Audio Signals Using Perceptual Noise Criteria

JAMES D. JOHNSTON

Abstract—A 4 bit/sample transform coder is designed using a psychoacoustically derived noise masking threshold based on the short-term input spectrum of the signal. The coder has been tested in a formal subjective test involving a wide selection of monophonic audio inputs. The signals used in the test were of 15 kHz bandwidth, sampled at 32 kHz. The bit rate of the resulting coder was 128 kbits/s. The subjective test shows that the coded signal could not be distinguished from the original at that bit rate. Subsequent informal work suggests that a bit rate of 96 kbits/s may maintain transparency for the set of inputs used in the test.

I. Introduction

CURRENT storage requirements for monophonic digital audio signals are about 705 kbits/s (16 bits * 44.1 kHz). Digital transmission facilities, to the extent that they exist, do not have the bandwidth or performance (either dynamic range or quality) normally considered to be high fidelity. In the studio, on a Compact Disc (CD), or the recent Digital Audio Tape (DAT), the storage requirements are met by the current technology of 16 bit PCM recording. While transmission at the bit rate for standard 16 bit PCM is feasible for some applications, current work that estimates the perceptual entropy for 15 kHz bandwidth signals suggests that something in the neighborhood of 2 bits/sample (64 kbits/s for 15 kHz bandwidth, or 88 kbits/s for 20 kHz bandwidth) [6] is sufficient for an efficient transparent coding scheme. This bound on bit rate for transparent coding has been named "Perceptual Entropy" (PE).

Digital transmission of audio signals, if available with the requisite performance and performance and affordability, could be used in a number of places, such as for remote broadcast lines, studio links, satellite transmission of high-quality audio, and the like. For instance, a channel with 15 kHz bandwidth that is perceptually transparent would be very useful to FM broadcast media, who could use it to provide a remote link free of the noise and frequency response problems normally associated with leased and/or equalized lines.

Previous work on digital audio has been concerned with relatively simple operations for coding and decoding, with low or moderate amounts of coding delay [3], [5], [9], [8], [1]. For commentary grade (7 kHz bandwidth) audio, the techniques have employed various combinations of subband coding and differential PCM, generally at a bit rate of 4 bits/sample. While the best examples of these algorithms can provide good teleconference quality coding of 7 kHz speech, transparent quality over a wide class of music signals has not been demonstrated.

Work for wider bandwidth signals, of 15 or 20 kHz, has centered primarily on simple bit rate reduction techniques that operate in the neighborhood of 6–10 bits/sample for near-transparent coding [11], [10]. The perceptual problems with the wider bandwidth are even more difficult than those at 7 kHz, due to the greater spectral tilt, spectral complexity, and dynamic range requirements.

The coder described in this paper uses a longer encoding delay and has considerably greater complexity, but provides transparent coding of a wide class of audio signals at 4 or even 3 bits/sample. In particular, it has been shown, by subjective testing, to provide transparent encoding of 15 kHz bandwidth audio at a bit rate of 4 bits/sample, or 128 kbits/s. This coder is intended for channels or storage media that provide good digital service, with an effective error rate of at least $5.7 * 10^{-6}$ and is not optimized, or intended, to be used for easily corrupted digital channels. If this coder is to be used with channels that have a high effective error rate, some parts of the encoded data must be provided with additional error correction. Furthermore, because of the delay, it is assumed that echos do not exist in the channel, or that half-duplex transmission is used.

The coders from [3] and the other references above have used source models as the primary basis for gains in coding efficiency. These coders provide either flat noise floors, or noise floors shaped according to simple statistics of the source model. In addition, some of the coders have used source-analysis methods such as backward adaptive predictors that have adaptation time constants that are slower than the variation in music signals [7].

Rather than base the coding algorithm on source models, the new coder, called the Perceptual Transform Coder, or PXFM for short, uses a human auditory model to derive a short-term spectral masking curve that is directly implemented in a transform coder. The redundancy in the signal is extracted only as is implicit in the frequency analysis and masking curve. Because unmasked noise in atypical input spectra [3], [7] has been shown to be a primary cause of the perceptual problems, the coder described below shows better perceptual performance than

Manuscript received May 31, 1987; revised October 19, 1987.
The author is with AT&T Bell Laboratories, Murray Hill, NJ 07974.
IEEE Log Number 8718539.

0733-8716/88/0200-0314$01.00 © 1988 IEEE

Exhibit 9 Page 363

published coders with signal-redundancy based source models that operate at the same bit rate, even though the signal-to-noise ratio of the perceptually coded signal may be remarkably low. Due to the method of auditory analysis, the problems of stationarity and spectral short-term fit demonstrated in [7] are bypassed.

Section II of the paper discusses the PE of 15 kHz wideband music signals. Section III discusses ways to implement the noise masking information that the PE calculation provides. Section IV provides the actual details of the PXFM coder. Section V describes the design, conditions, and results of a listening test using PXFM coder at 4 bits/sample, with a 32 kHz sampling frequency on 15 kHz audio signals. Sections VI and VII wrap up the paper and briefly describe ongoing work.

II. Perceptual Entropy and Perceptual Threshold

As mentioned above, [6] discusses the perceptual entropy of various audio signals. The PE is calculated by estimating a threshold that represents the maximum level of injected noise that will be inaudible when added to the input signal.

A. *Perceptual Entropy of 15 kHz Bandwidth Signals*

The perceptual entropy estimates mentioned in the Introduction are short-term (64 ms) parameters of any arbitrary audio signal. A histogram of the PE of the entire set of sources is shown in Fig. 1. Table I is a list of the source material in Fig. 1. These entropy measures indeed suggest that coding at rates well less than 16 bits/sample should be feasible for coding of audio signals. In the example shown, the PE of 1.8 percent of the signals is above 2 bits/sample, and none is above 2.23 bits/sample.

A breakdown of the PE by (selected) source is shown in Fig. 2. It is interesting that many signals considered "hard" to encode, such as piano and flute, do not have a high entropy measure. On the other hand, signals such as a capella vocals and rock music have a very high entropy measure, even though they are not usually considered as difficult or sensitive to distortion.

B. *The Perceptual Threshold*

In the PE calculation, a noise threshold is calculated for each time segment of the source material. This threshold, which estimates the maximum noise that can be inaudibly inserted into the signal, can also be used in a coder to maximize the amount of perceptually acceptable noise that can be injected into a given signal segment. In an actual transform coder, the direct use of this threshold would result in a variable bit rate coding scheme that would not be well matched to most transmission or storage systems, so an actual coder must have an algorithm that varies the threshold appropriately. The mechanism for adapting the coder to a given bit rate will be discussed below.

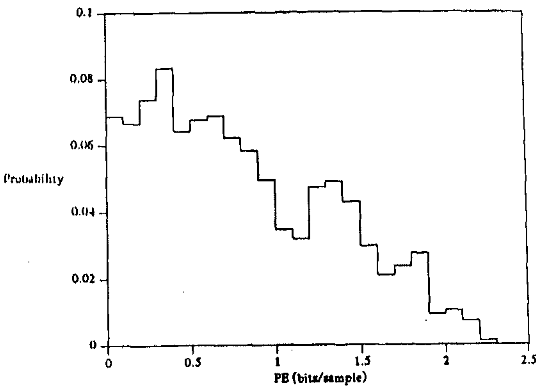

Fig. 1. A histogram of the PE for all sources listed in Table 1.

TABLE I
A List of the Source Materials Used in Both the PE Study and the PXFM Listening Test

15kHz Source Material, sampled at 32kHz	
4 Letter Code	Instrument/Style
ATBR	Brass Choir, baroque music
BJPV	Pop Vocal, 50's style
BPSP	Solo Piano (soft passage)
DZFV	Female Coloratura Opera Vocal
DZMV	Male Operatic Bass Vocal
EHPV	Female Pop Vocal
JBHC	Complex Harpsichord Solo
JBOR	Baroque Organ
LFSR	Solo Recorder
LZBD	"Heavy Metal" Rock Vocal/Instrumental
PEAK	Whip-crack (surrounded by silence)
PERC	Percussion (African Drums)
PHOR	Romantic Style Full Organ
WHEP	Electronic Pop/Bell Synthesizer
WHSG	Solo Guitar

III. Why a Transform Coder Is Used

The primary goal in using the perceptual information is to implement the noise threshold. While the noise threshold shape could be used in other types of coders, a transform coder represents the most direct way to implement a particular noise shape accurately. The use of the noise threshold as a noise-shaping parameter in a subband or adaptive predictor DPCM coder is harder to implement directly.

For instance, in a DPCM coder, the implementation of the noise threshold shape would require that a sufficiently accurate model of the noise shape be made in LPC parameters. Due to the nonuniform band structure in frequency and the number of different coefficients that must be fit (14 for telephone speech, 26 for music, where the number of coefficients that must be fit depends on the number of critical bands in the signal bandwidth), the order of a noise-shaping filter would be between 28 and 56. This noise-shaping filter would have to be recalculated at least each 64 ms, and parameters would have to be interleaved with the data in order for the receiver to be able to reconstruct the signal. Additionally, a method of slowly changing the noise shaping and LPC gain parameters, taking the noise shaping and LPC gain trajectory carefully into account, would have to be developed.

The critical band frequency divisions do not lend themselves conveniently to subband coding, either of the standard tree structure [4] or the multiband structure [2]. As-

Exhibit Q Page 364

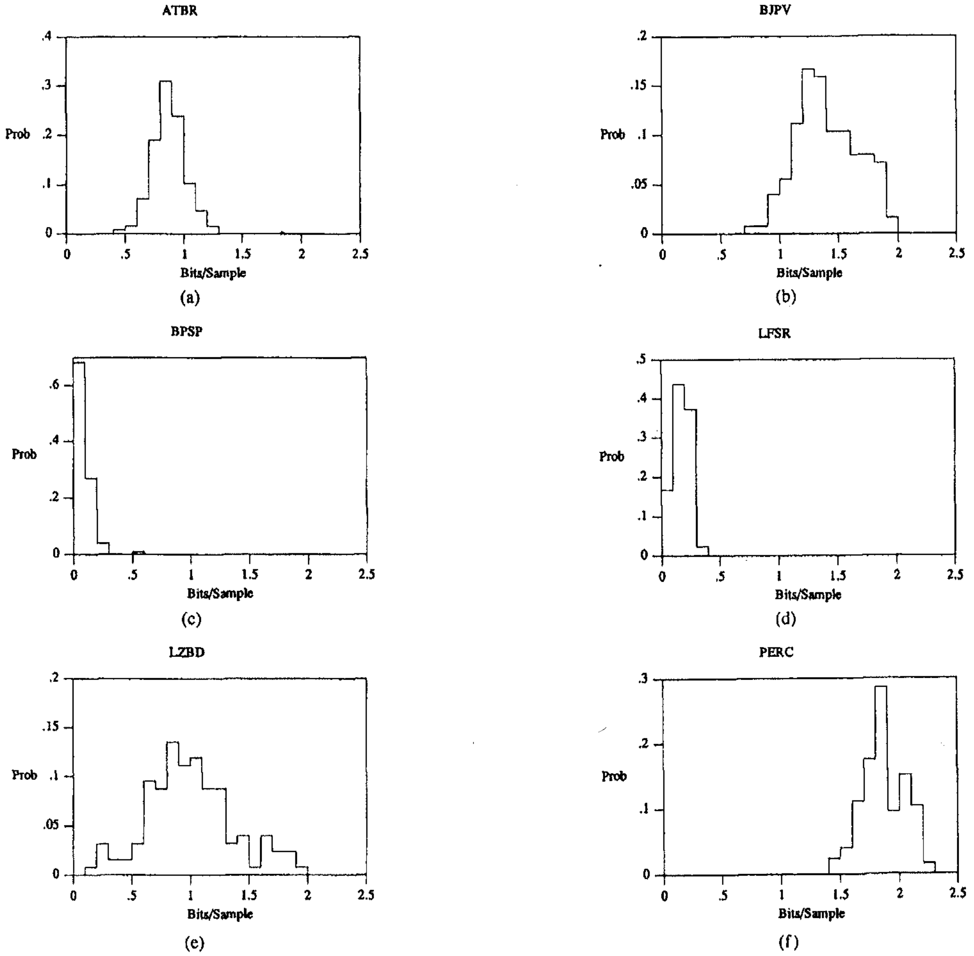

Fig. 2. (a)-(f) Individual histograms of PE for some of the sources listed in Table 1.

suming that an appropriate set of subbands can be designed, a bit allocation algorithm could be devised to enforce the noise spectra. An algorithm implemented in this way would have to be carefully designed to account for dynamic behavior on musical attacks, drumbeats, etc., as well as encode sinusoidal signals in an optimal fashion in order to provide efficient encoding.

This particular transform coder, with side information included and no memory beyond the current block of encoding, provides a way to provide the noise shaping, while maintaining instant response to amplitude changes, etc. The question of transient response becomes one of bandwidth, i.e., one must encode the wider bandwidth that results from rapid transitions in time and consequently use a higher bit rate.

The side information in this particular coder includes the quantized threshold levels and quantized spectral peak levels. This information allows the receiver to recover the bit allocation in the same way as the transmitter, thus making explicit information on bit allocation unnecessary. Since there is no block-to-block memory in the coder, other than overlap-add, the handling of rapid time transition problems, such as peak clipping and slope overload, is facilitated.

IV. Particulars of the Perceptual Transform Coder

A block diagram of the transform coder, as implemented in software, is shown in Fig. 3. The input signal is windowed and processed by a 2048 point real-complex FFT. The spectrum is used both as the source of the transform coder quantizers and for the perceptual threshold calculation. As mentioned above, the perceptual threshold is used as the noise-shaping function for the coder. The quantization step sizes for the frequency components are determined by the bit-rate calculation process, which maintains the shape of the perceptual threshold. The

Exhibit Q Page 365

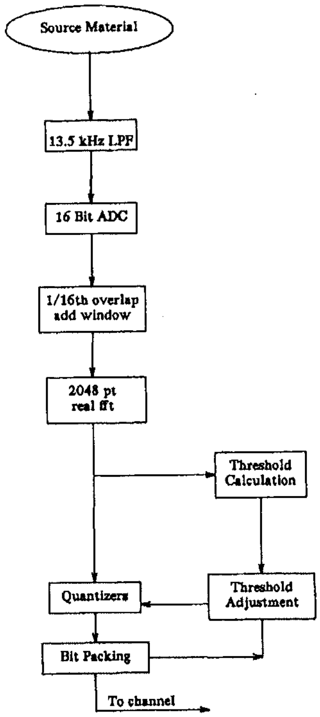

Fig. 3. Block diagram of the PXFM coder.

quantizer step sizes are calculated iteratively to provide a fixed bit-rate coder.

After the quantizer step sizes are calculated, the quantized frequency components are calculated and compressed for transmission along with the side information from the bit-rate adjustment process.

The windowing process for overlap-add and analysis provides for an overlap-add of 1/16th. The window that is used in the overlap sections is the square root of a Hanning window of length equal to twice the total overlap. The total number of data points is 2048. With the overlap of 1/16th, the number of new data points processed in each block is 1920.

The FFT is implemented directly on the windowed data. The 2048 real input data are converted to a spectrum of 1024 complex points, counting the dc and Nyquist frequency points as one complex point.

After the initial threshold is calculated, the bit rate, including the compression/bit-packing algorithm, is calculated, using the FFT data and the perceptual threshold. The threshold, a vector of 26 scalars, each scalar corresponding to one critical band (there are 26 critical bands in the 15 kHz bandwidth), is multiplied by an estimator in a search procedure until the bit-rate limits are met. While the upper bit rate cannot be exceeded, the lower bit rate may not be met by very low energy signals. The bit-rate calculation is designed such that very low energy signals will be overquantized by a maximum of 20 dB. During the bit-rate calculation process, side information regarding the peak levels of the spectrum in a set of 128 fixed bands is calculated and quantized as side information.

After the threshold is calculated for the particular data block, the quantized threshold values and spectral peak values are available for transmission. At this point, the quantized FFT data are calculated, concatenated to the side information, and provided to the decoder.

The perceptual threshold calculation, spectral peak calculation, side information description, bit-packing algorithm, and the bit-rate calculation algorithm are detailed below.

A. Calculation of the Masking Threshold

There are several steps involved in calculating the masking threshold. They are

- critical band analysis of the signal
- applying the spreading function to the critical band spectrum
- calculating the spread masking threshold
- accounting for absolute thresholds
- relating the spread masking threshold to the critical band masking threshold.

1) Critical Band Analysis: We are presented with the spectrum $\mathrm{Re}(\omega)$, $\mathrm{Im}(\omega)$ of the signal from the FFT. The complex spectrum is converted to the power spectrum,

$$P(\omega) = \mathrm{Re}^2(\omega) + \mathrm{Im}^2(\omega).$$

The spectrum is then partitioned into critical bands, according to [17], see Table II, and the energy in each critical band summed, i.e.,

$$B_i = \sum_{\omega = bl_i}^{bh_i} P(\omega)$$

where bl_i is the lower boundary of critical band i, bh_i is the upper boundary of critical band i, and B_i is the energy in critical band i, where $i = 1$ to $i_{\max}$, and $i_{\max}$ is dependent on the sampling rate. Fig. 4(a) shows a power spectrum and critical band spectrum for 64 ms of a loud brass passage.

A true critical band analysis would sum across one critical band at each ω in order to create a continuous critical band spectrum. For the purposes of the PE calculation, the discrete critical band represents a close approximation. B_i is then passed to the spreading function.

2) Spreading Function: The masking estimates from [17] and [18] provide information on masking of signals by signals within the same critical band. The spreading function as given in [16] is used to estimate the effects of masking across critical bands. The spreading function is calculated for $abs(j - i) \leq 25$, where i is the bark frequency of the masked signal, and j is the bark frequency of the masking signal, and placed into a matrix S_{ij}. The term bark is often used to indicate a frequency difference of 1 critical band. The convolution of the $B(\omega)$ with the spreading function is implemented as a matrix multipli-

Exhibit Q Page 366

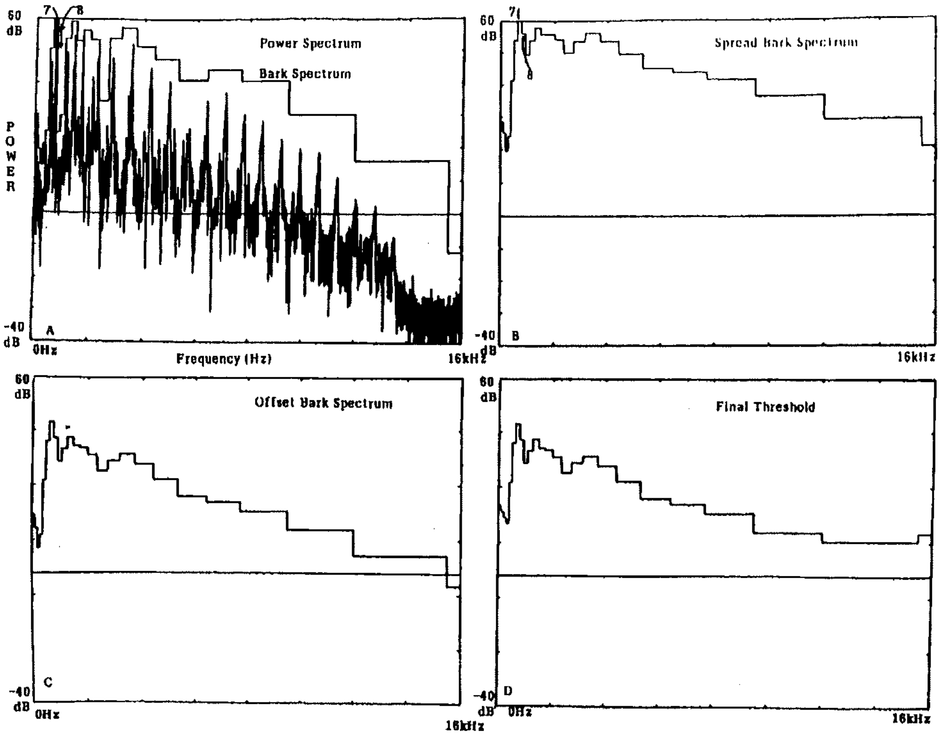

Fig. 4. A graphic example of the calculation of the perceptual threshold for a section of trumpet (ATBR) music.

TABLE II
A LIST OF CRITICAL BAND CENTER AND EDGE FREQUENCIES AS USED IN THIS PAPER

Table of Critical bands, from Scharf, et. al.			
Bånd number Hz	Lower Edge Hz	Center Hz	Upper Edge Hz
1	0	50	100
2	100	150	200
3	200	250	300
4	300	350	400
5	400	450	510
6	510	570	630
7	630	700	770
8	770	840	920
9	920	1000	1080
10	1080	1170	1270
11	1270	1370	1480
12	1480	1600	1720
13	1720	1850	2000
14	2000	2150	2320
15	2320	2500	2700
16	2700	2900	3150
17	3150	3400	3700
18	3700	4000	4400
19	4400	4800	5300
20	5300	5800	6400
21	6400	7000	7700
22	7700	8500	9500
23	9500	10500	12000
24	12000	13500	15500
25	15500	19500	

cation, i.e., $C_i = S_{ij} * B_i$. The value of C_i denotes the spread critical band spectrum. Fig. 4(b) shows the results of spreading of the bark spectrum in Fig. 4(a).

3) Calculating the Noise Masking Threshold: There are two noise masking thresholds detailed in [12]–[18] and discussed in the Appendix. The first, for tone masking noise, is estimated as $14.5 + i$ dB below C_i, where i is the bark frequency, where this estimate is from [17] via [16]. The second, for noise masking a tone, is estimated as 5.5 dB below C_i uniformly across the critical band spectrum. The estimate for noise masking of tones is due to [18].

In order to determine the noiselike or tonelike nature of the signal, the Spectral Flatness Measure (SFM) is used. The SFM is defined as the ratio of the geometric mean (Gm) of the power spectrum to the arithmetic mean (Am) of the power spectrum. In this use, the SFM is converted to decibels, i.e.,

$$SFM_{dB} = 10 \log_{10} \frac{Gm}{Am},$$

and further used to generate a coefficient of tonality α as follows:

$$\alpha = \min \left(\frac{SFM_{dB}}{SFM_{dBmax}}, 1 \right),$$

i.e., an SFM of $SFM_{dBmax} = -60$ dB is used to estimate that the signal is entirely tonelike, and an SFM of 0 dB to indicate a signal that is completely noiselike. In other words, an SFM of -30 dB would result in $\alpha = 0.5$, and an SFM of -75 dB would result in $\alpha = 1$.

The offset (O_i) in decibels for the masking energy in

Exhibit Q Page 367

each band i is then set as

$$O_i = \alpha(14.5 + i) + (1 - \alpha)5.5.$$

In other words, the index α is used to weight geometrically the two threshold offsets, $14.5 + i$ dB for tone masking noise and 5.5 dB for noise masking tones.

The threshold offset is then subtracted from the spread critical band spectrum to yield the spread threshold estimate T_i

$$T_i = 10^{\log_{10}(C_i) - (O_i/10)}.$$

In practice, the use of the SFM to estimate the tonality of signals is useful, as most tonelike signals such as organ, sine waves, flute, etc., have an SFM that is close to or over the limit, and signals such as percussion have SFM's, in transient sections, that are between −5 and − 15 dB. Speech signals of 200–3200 Hz bandwidth are in the range of −20 to − 30 dB. Fig. 4(c) shows the plot of the spread threshold estimate for the data in Fig. 4(a) and (b).

4) Converting the Spread Threshold Back to the Bark Domain: Strictly speaking, the convolution of the spreading function with B_i must be undone, i.e., the threshold calculated as T_i should be deconvolved. This process is very unstable due to the shape of the spreading function, and often leads to artifacts such as a negative energy for a threshold, zero thresholds, etc. The unusual errors come about because the deconvolution process seeks a strictly numerical solution that disregards the physical and acoustic realities of the situation.

In place of the deconvolution, renormalization is used. The spreading function, because of its shape, increases the energy estimates in each band due to the effects of spreading. The renormalization takes this into account, and multiplies each T_i by the inverse of the energy gain, assuming a uniform energy of 1 in each band. In other words, given a flat B_i, and a condition where all O_i are equal, it will return a flat renormalized T_i. This renormalized T_i will be denoted T_i'.

While this particular calculation is not optimum, calculating the optimum would require a max-min sort of optimization of the deconvolution, with bounds on the acceptable renormalized T_i, that would search out a threshold that fits both the deconvolution and provides a minimum in bit rate. This process is not easily or efficiently implemented, and accounts for very little of the error in the bit-rate estimation process.

Including Absolute Threshold Information: After the noise energy is renormalized in the bark domain, the bark thresholds are compared to the absolute threshold measurements due to [12]. Since the masking thresholds have thus far been calculated without reference to absolute level, they must be checked to make sure that they do not demand a level of noise below the absolute limits of bearing.

The absolute thresholds are set such that a signal of 4 kHz, with a peak magnitude of ±1 least significant bit in a 16 bit integer is at the absolute threshold of hearing. Any critical band that has a calculated noise threshold lower than the absolute threshold is changed to the absolute threshold for that critical band. At high and low frequencies, the absolute threshold varies inside the critical band. In such cases, the mean of the critical band edges is used.

Fig. 4(d) plots the final threshold, after renormalization and adjusting for absolute threshold conditions. This is the threshold used to start the bit-rate adjustment procedure. This threshold is denoted Th_j.

B. Spectral Peak Calculation

The spectral peaks of each 8 complex points of the transform, P_i are calculated as follows:

$$Sr(\omega) = abs(\mathrm{Re}(\omega))$$
$$Si(\omega) = abs(\mathrm{Im}(\omega))$$
$$P_i = \max\left(Sr(8*(i-1)+1:8*i),\ Si(8*(i-1)+1:8*i)\right)$$

where Sr and Si are temporary storage for the absolute values of the spectrum, and the notation $Sr\,(a:b)$ means that the values between a and b of Sr are used. Since the particular transform is a real-complex transform of length 2048, there are 1024 unique complex pairs, and 128 spectral peak values. These spectral peak values are quantized in a log format, using 8 bits, where each step in the quantizer represents $\frac{170}{256} = 0.664$ dB. The number 170 is explained below.

The choice of 8 complex lines per spectral peak was made by trying several different values, particularly 4, 8, 16, and 32, and using the one that had the best overall bit rate for a given quantization accuracy. The tradeoff is between bits used for side information versus bits used for empty quantizer slots.

The size of the quantizer is determined by the possible dynamic range of the 2048 point spectrum, given a 16 bit PCM input source. In a transform, the total spectral dynamic range can vary from that of a single unit impulse, here set to quantizer level 1, to the spectral energy of a full scale sinewave, roughly 167 dB higher, hence the total range of 170 dB.

C. Side Information Description

There are three parts to the side information. The first is a 16 bit PCM encoding of the dc term from the FFT. The second is the 25 critical band threshold levels, as modified by the bit-rate calculation algorithm. These 25 terms are encoded in 8 bit log PCM, where each step in the quantizer represents $\frac{170}{256} = 0.664$ dB. The quantized threshold levels are used in the bit-rate calculation process. The third part of the side information is a vector of 128 quantized peak levels.

The side information is used in both the transmitter and receiver to calculate the bit allocation. In the transmitter, the side information is also used for bit-rate calculation.

Exhibit Q Page 368

D. Bit-Rate Calculation

Given the quantized spectral peaks and quantized initial threshold, Th_j, as well as the original spectrum, it is possible to calculate the number of bits that are required to encode the block of the signal, given the current threshold. On the first pass through bit-rate calculations, $Thr_j = Th_j$. The bit-rate adjustment procedure will be applied to Thr_j. There are two steps.

- Calculation of the number of levels in each set of quantizers, where a set of quantizers is determined by the peak level measurements.
- Packing of the quantized numbers. Since the algorithm may use a simple compression algorithm to aid in bitpacking, the length of the packed data may not be immediately calculable from the $\Sigma \log_2(k)$ of the number of levels k in each set of quantizers.

The number of levels in each set of quantizers is determined by calculating $k_i = 2 * nint(P_i / Thr_j) + 1$, where k_i is the number of levels in each quantizer, P_i is the quantized peak level, Thr_j is the appropriate quantized threshold value, and *nint* represents the "nearest integer" function.

1) Bit-Packing Algorithm: There are two forms of bit packing that can be used. The simpler form uses a variable radix technique that works as follows.

- First, using a deterministic sort, sort all of the k_i's and their corresponding quantizer values in decreasing order.
- Second, fill a 64 bit word in steps, at each step doing variable radix arithmetic to encode the largest radix that will still fit in the word.
- Continue this process with 64 bit words, until all the data are expanded.

This bit-packing algorithm wastes a small amount, usually about $\frac{1}{64}$, of the bit rate allocated for data. This loss can be accounted for in the initial bit-rate calculation, making the second step of bit-rate calculation, calculation of the compressed data size, unnecessary.

The more complicated bit-packing algorithm uses a modified Huffman code with performance that varies in a data-dependent manner. If this compression method is used, the compressed bit rate must be calculated from the quantizer data as well as from the number of levels in each quantizer.

The current performance of the modified Huffman code scheme, using a fixed codebook, is between a compression rate of 0.8 and 1.2, where 1.2 constitutes a 20 percent expansion in the size of the data. As the bit rate of the coder is reduced, the Huffman coding becomes a bit more efficient, providing compression rates between 0.75 and 1.1.

Research into the use of a Huffman code as a compression/bit-packing algorithm is continuing. Currently, the results of the coder using the radix bit-packing scheme are reported.

E. Bit-Rate Adjustment

The bit-rate adjustment is done by multiplying the Thr_j's by an estimator. The estimator is selected by the past history of the coded signal and by the amount by which the bit rate is above or below the intended bit rate. This process is repeated, with a temporary copy of the estimator raised to the $\frac{1}{3}$ power whenever the signal of the error changes. The nonquantized thresholds Thr_j are the thresholds that are modified, although the quantized thresholds are used for bit-rate determination.

F. After Bit-Rate Adjustment

When the bit rate comes within the set limits, which depend on the bit rate desired for the coder, the data are compressed/packed, combined with the side information, and transmitted. The data, with a block size 1920 times the rate in bits per sample, are then passed to the receiver.

G. The Receiver

The receiver unpacks the bits, using the same bit-packing scheme as the transmitter, and reconstructs the spectrum. The reconstructed spectrum is passed through an inverse transform, and the results windowed with the same window used for analysis, and added to the previous overlap section. The first 1920 samples are then passed to the data file or analog reconstruction device.

The quantized P_i are required at the receiver in order to know the unpacking radixes. The Thr_j are used as the step sizes of the unpacked data. The receiver calculates the order of data received by doing the same sort and packing of k_i's.

V. Subjective Testing of the PXFM Coder

There were two questions that we wanted to resolve with listener evaluation. The first was the question of window length, i.e., which block size, of 32, 64, or 128 ms, was preferred? The second was the question of quality, i.e, could the listener distinguish between the original and most preferred coder from the first test?

In addition, we wanted to discover what sort of listening experience and/or listening ability would lead to accurate detection of the coded signals. For that purpose, a prescreening test, using tone masking of critical band noise, as in [17], was also designed. This test, along with a questionnaire on listening habits, noise exposure, and auditory health and knowledge, was given to each subject.

A. Source Materials for Testing

We used a total of 14 source materials for the subjective test. All were music signals, taken from compact disc recordings where the two channels were summed to a monophonic signal. The monophonic signal was low pass filtered at 15 kHz, and sampled at 32 kHz, with 16 bit resolution. The gain of the signal path was fixed over all signals, such that the peaks of the loudest signal were digitized to an amplitude of > ±32 000. No overload was allowed. The material varied in composition from hard rock to baroque harpsichord. Table I lists the 14 sources. These source materials are the same ones whose perceptual entropy is mentioned above. The 14 sources were used in both experiments. The sine wave and critical band

Exhibit Q Page 309

noise for the screening test were digitally generated with 32 bit accuracy, and quantized with dithering to 16 bits via computer simulation.

B. Encoding Methods for the Coded Signals

For the first test, each signal was encoded in software simulation by the PXFM coder. There were three encodings of each signal, at window lengths of 32, 64, and 128 ms. In each case, the bit rate of the encoding was set to 4 bits/sample. With 4 stimuli per signal, original plus three coded, 14 * 4 = 56 stimuli were used in the actual test.

In the second test, the 64 ms coder was selected (see below). All of the sources were encoded by this coder, using the same methods, for a total of 14 * 2 = 28 stimuli.

C. Selection of Subjects

Experience with pilot studies involving high-quality audio have shown that listeners with experience, i.e., musicians, audio enthusiasts, concertgoers, recording engineers, and the like are more likely to make consistent judgments than inexperienced subjects, so our largest set of subjects consisted of individuals with some listening experience or interest.

There were three main groups used as subjects. They were as follows:

- AT&T Summer Research Program employees. These people are typically age 19–22, roughly evenly divided male/female, and have a wide range of musical ability, listening habits, and musical/audio interests.
- AT&T–BL speech researchers. These subjects are not to be considered innocent subjects, since the majority of these subjects were taken from the Speech Processing, Signal Processing, and Acoustics Research departments. While a wide range of outside interests and listening habits prevailed, these subjects are involved in either signal processing or coding on a professional basis. These subjects happened to be primarily male.
- AT&T–BL Audiophile Club members. These people are all interested in high-quality sound reproduction. They typically listen to a considerable amount of recorded music in their own home. Their professions vary widely, from researcher to painter. The music preferences of each of these subjects is usually quite eclectic, most including classical/romantic, modern, rock, jazz, baroque, operatic, and more.

Each subject was asked in the questionnaire if they considered themselves "an experienced listener;" 35 of 57 subjects answered "yes," and 22 answered "no" or "maybe."

D. Signal Format

For each of the tests, prescreening included, the format of a trial was:

- 2 s of stimulus *A*
- 1 s of silence
- 2 s of stimulus *B*
- 7 s of silence.

In any one trial, the two stimuli differed only in coder type, the source material was the same. The 2-s segments taken from each stimulus were windowed with a 256 point raised cosine at the beginning and end to prevent artifacts at onset and end of the segments.

E. The Actual Test Recordings

There were three sets of actual Sony 16 bit PCM-F1 digital recordings used in the test, one for the screening test, one for the window length section, and one for the transparency testing.

1) The First Test Tape: For the first test, which compares 4 different types of stimuli for each kind of source material, a complete test of pairs was done, i.e., 12 pairs were presented to the subject in the format mentioned above. The pairs of identical material were not presented.

The order within each set of source material was randomly determined by computer, and all of each set of source material was presented at one time. The pairs of each source were presented in a block.

The starting points of the excerpts of each signal were chosen randomly, with the constraint that 2 s must remain in the stimulus. The same starting point was used for both parts of each pair of stimulus. For each starting point, both orders were used, i.e., 6 different starting points were chosen, and each starting point used twice, once for each order.

The sequence and starting points that were initially chosen were presented to all subjects. The materials were reproduced by 16 D/A conversion, and recorded on a Sony PCM-F1 digital recorder, in 16 bit mode. This reproduction and recording method was used for all tests detailed in this paper.

2) The Second Test Tape: In the second test, the choice was presented between an uncoded signal and a coded signal. For this test, 4 starting points were randomly selected with each source material and used in both orders, for a total of 8 pairs for each source material.

In this test, two instances of the signal for the Romantic Pipe Organ selection were corrupted. The corrupting signal was a uniform, uncorrelated random number stream of $\sigma = 1$ and $\mu = 0$. This number stream was added to each sample, with an additional scaling factor of 0.001 * *abs* (*samplevalue*). Essentially, this adds noise 60 dB below the local energy present in the signal. The position of these 4 corrupted trials was in the next to last set of 8 trials in order not to provide an "anchor" for the test.

3) The Screening Tape: The screening tape compared a clean (to 90 dB or better) sine wave to a sine wave with critical band noise added. The level of the added noise was set to 7 levels, centered on 24.5 dB below the energy of the sine wave, and the level was varied ± 3, ± 6, and ± 9 dB in either direction.

Each level of noise was compared only to the clean sine wave. The signal was presented in 4 sets of 14 pairs, where each set of 14 pairs had each comparison in both orders. As above, the order was selected randomly.

Exhibit Q Page 370

F. Audio Reproduction Facilities

The output from the Sony PCM-F1 was sent to a set of Stax-Mk III Pro headphones driven by a Stax SRM-I/Mk-II amplifier. The signal was applied to both channels of the headphones. The stimuli were presented in a single-wall sound booth with an A-weighted SPL of 20–25 dB in the interior when the subject was not present. All presentation stimuli were monitored by the test administrator at all times. The presentation level was adjusted such that the peak level of the loudest signal was about 95 dB SPL.

G. The First Test

The test procedure for the first test, between different coder window lengths, was as follows.

- The subject was asked to complete a questionnaire that asked about age, sex, listening experience, musical training, listening inclination, noise exposure, and prior ear testing.
- The screening signals were presented in 4 groups of 14 signals, as above. The subjects were told that one of each pair of signals was a pure tone, the other a noisy tone. The task was to select the pure tone of each pair.
- After the tone test, the subject was given a 5 min break.
- The subject was given instructions that explained that each pair contained signals that were processed in some fashion. The task was to select the signal that the subject preferred.
- The subject was allowed a 5 min break after each set of four source materials, if desired.
- After the last group of 2 source materials was completed, the subjects were asked if they had any comments.

The test administrator was not aware of the ordering information while the test was being presented. The total time required for the test was approximately 50–60 min, including rest breaks. Twenty subjects were tested for this group. The responses of one subject were discarded when that subject was observed flipping a coin for 2/3 of the music section of the test.

1) Results: In this test, all sets of comparisons were at chance levels. The following table shows the preferences. While none of the scores is significant, the best performing coder, the one with the 64 ms window, was selected for the second test. There were no detectable order effects in this test, so the results are combined for both orders.

The number in any location in the table shows the probability of the signal indicated on the row of the table being selected over the coder indicated in the column.

	Original	128 ms	64 ms	32 ms
Original	—	0.54	0.46	0.51
128 ms	—	—	0.47	0.50
64 ms	—	—	—	0.52

The results of the tone tests for these 19 subjects were consistent with earlier tone-masking-noise tests, as detailed in [17]. Several of the subjects, who showed a tone discrimination much better than average, were invited back for the second set of tests.

H. The Second Test

The procedure for the second test, which attempted to discover listener discrimination between the encoded source and the original signal, was as follows.

- The subject was asked to complete a questionnaire that asked about age, sex, listening experience, musical training, listening inclination, noise exposure, and prior ear testing. The 5 subjects who were recruited from the first test were not given the questionnaire again.
- The screening signals were presented in 4 groups of 14 signals, as above. The subjects were told that one of each pair of signals was a pure tone, the other a noisy tone. The task was to select the pure tone of each pair. Again, the 5 subjects who were recruited from the first test were not given the tone test.
- After the tone test, the subject was given a 5 min break.
- In the coder test, the subject was given instructions that explained that each pair contained one processed and one original signal. The task was to select the original signal.
- The subject was allowed a 5 min break after each set of four source material groupings, if desired.
- After the last group of 2 source materials was completed, the subjects were asked if they had any comments.

The test administrator was not aware of the ordering information while the test was being presented. The total time required for the test was approximately 45–55 min, including rest breaks. Forty-three subjects were tested for this group. No subjects were excluded from the final data analysis.

1) Results: In this test as well, the final results show that the subjects could not detect the difference between stimuli, except for the corrupted signal. For the corrupted signal trials, the probability of selecting the clear signal was 0.87. Many of the subjects commented (or complained) that "they all sound the same." A few subjects were sure that they had discovered differences in some source materials, but their data indicate that they were selecting randomly.

A few of the subjects noted, either to the administrator or on the answer sheet, that one or two of the pipe organ signals sounded "scratchy" or "noisy." Even though a few people noted this or mentioned it to the test administrator, nearly all subjects discriminated strongly against the corrupted pipe organ signal.

The final selection probability, overall (not including the corrupted signal), was 0.47, which represents the probability of selecting the original signal. Selection by source material or by subject is similarly neutral, suggesting that the 4 bit coder is indistinguishable from the original to the limits of this particular test. The test is shown by the detection of the computed signal to be at least reasonably sensitive.

Exhibit Q Page 371

VI. Conclusions

There are two main conclusions to be reached from this work. The first is that the current methods of digital encoding are very inefficient in the informational sense. The second is that the coder described in this paper, when used at 4 bits/sample in a 32 kHz sampling rate system, is more than adequate for transparent coding.

More recent work, and some informal testing, suggest that the same coder used in this listener evaluation can be used at a bit rate of 3 bits/sample for transparent coding of the same sources. This informal listening is corroborated by the PE statistics, which show no signals that exceed 2.23 bits/sample, and very few above 2 bits/sample, suggesting that a bit rate of 4 bits/sample constitutes significant overdesign for this coder. The coder itself has been modified to record its estimate of the number of bits necessary for transparency, using the same threshold used in PE measurement, and taking into account the coding inefficiencies. This measure peaks at about 3.1 bits/sample, and in general is well below 2 bits/sample, supporting the informal results that suggest that 3 bits/sample is either transparent or nearly transparent. Work relating to increased coding efficiency, efficient stereo encoding, and more effective compression/packing algorithms is underway.

Examination of some 20 kHz bandwidth signals sampled at 44.1 kHz suggest that the same or slightly fewer bits/sample should provide transparent coding. The lower energy at very high frequencies requires fewer than average bits for the additional bandwidth, resulting in the lower rate measured in bits/sample.

A. Coding in the Presence of Channel Errors

Because this coder extracts a great deal of redundancy, it is likely to be sensitive to channel errors. For some uses, where hardwired lines or digital storage media are used, the sensitivity to errors is not likely to be a problem. Where channel errors are worse than approximately $5.7 * 10^{-6}$, some protection of the side information must be done in order to prevent incorrect calculation of the bit allocation.

Appendix

A discussion of critical bands can be found in *Foundations of Modern Auditory Theory* (New York: Academic, 1970) edited by J. V. Tobias. In particular, chapter 5, by B. Scharf, introduces the idea of critical bands, and provides some feeling for the subject.

A short historical discussion, with many references, appears in *Handbook of Perception, Volume IV*, edited by E. C. Carterette and M. P. Friedman, on pp. 172 and 173 in Whitfield's chapter on "The Neural Code."

Acknowledgment

I would like to acknowledge the essential help of J. L. Hall of AT&T Bell Labs in the psychoacoustic realm. In addition, I would like to thank N. S. Jayant for his technical advice and felicitous emendations, D. Bock for his help with real-time facilities and source material recording, S. Pruzansky for help with experiment design and documentation, D. Zuniga for her help in running and administering the listening test, and many others who bothered to answer or ask questions.

References

[1] CCITT Recommendation G.722, "7kHz audio coding with 64 kbits/s," working party XVIII/8, Geneva, 1987.

[2] R. V. Cox, "The design of uniformly and nonuniformly spaced pseudoquadrature mirror filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1090–1096, Oct. 1986.

[3] R. V. Cox, J. H. Snyder, R. E. Crochiere, D. E. Bock, and J. D. Johnson, "Testing of wideband digital coders," in *Proc. ICASSP*, 1984, pp. 19.3.1–19.3.4.

[4] D. Esteban and C. Galand, "Application of quadrature mirror filter banks to splitband voice coding schemes," in *Proc. ICASSP*, 1977, pp. 191–195.

[5] J. D. Johnston and R. E. Crochiere, "An all-digital 'commentary grade' subband coder," *J. Audio Eng. Soc.*, vol. 27, no. 11, pp. 855–865, Nov. 1979.

[6] J. Johnston, "A method of estimating the perceptual entropy of an audio signal," submitted to ICASSP '88.

[7] —, "Digital coding of musical sound—Some statistics of interest" (Abstract), presented at the IEEE ASSP 1986 Mohonk Conf. Audio.

[8] G. Modena, A. Coleman, P. Usai, and P. Coverdale, "Subjective performance evaluation of the 7 kHz audio coder," in *Globecom 1986 Proc.*, pp. 17.2.1–17.2.6.

[9] E. B. Richardson and N. S. Jayant, "Subband coding with adaptive prediction for 56 kbit/s audio," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 691–696, Aug. 1986.

[10] J. Soumagne, P. Mabilleau, S. Morissette, G. Chouinard, and D. Bennett, "A comparative study of the proposed high quality coding schemes for digital music," in *ICASSP 1986 Proc.*, pp. 1.6.1–1.6.4.

[11] C. Todd, "A digital audio system for broadcast and prerecorded media," presented at the 75th AES Conv. Paris, France, Mar. 1984.

[12] H. Fletcher, "Auditory patterns," *Rev. Modern Phys.*, vol. 12, pp. 47–65, 1940.

[13] R. Feldkeller and E. Zwicker, *Das Ohr als Nachrichtenempfanger.* Stuttgart: Hirzel, 1956.

[14] D. D. Greenwood, "Critical bandwidth and frequency coordinates of the basilar membrane." *J. Acoust. Soc. Amer.*, vol. 33, pp. 1344–1356, 1961.

[15] J. Zwislocki, "Analysis of some auditory characteristics," *Handbook of Mathmatical Psychology*, R. D. Luce, R. R. Bush, and E. Galanter, Eds. New York: Wiley, 1965.

[16] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, no. 6, pp. 1647–1651, Dec. 1979.

[17] B. Scharf, ch. 5 in *Foundations of Modern Auditory Theory.* New York: Academic, 1970.

[18] R. P. Hellman, "Asymmetry of masking between noise and tone," *Percept. and Psychophys.*, vol. 11, pp. 241–246, 1972.

James D. Johnston received the B.S.E.E. and M.S.E.E. degrees from Carnegie Mellon University, Pittsburgh, PA, in 1975 and 1976, respectively.

He has been employed at AT&T Bell Laboratories since 1976, first in the Acoustics Research Department, and then as a member of the Technical Staff in the Signal Processing Research Department. In the past, he has been involved in digital signal processing hardware and speech encoding research, focusing on algorithms that could be realized in real time. His current interests are very high quality speech and music coding at 6.5, 15, and 20 kHz bandwidths—especially those techniques that take advantage of the limits of the human ear, the measurement of bit-rate limits to transparent coding of signals that are to be presented to the human ear, methods of noiseless compression that can be combined with perceptually based coders, and estimation of masking thresholds for arbitrary audio signals. He contributes to publications of the IEEE and the Acoustical Society of America.

Mr. Johnston is a member of the Audio Engineering Society.

Exhibit Q Page 312